Results of the 2014 Survey of NSF-funded researchers on their bioinformatics (genomics) needs

Thomas G. Doak
William Barnett
Carrie Ganote
Le-Shin Wu

Indiana University

PTI Technical Report PTI-TR14-015

December 31, 2014

Citation:

Doak, T.G., Barnett, W., Ganote, C., & Wu, L.S. (2014) "Results of the 2014 survey of NSF-funded researchers on their bioinformatics (genomics) needs", Indiana University, Bloomington, IN. PTI Technical Report PTI-TR14-015. Retrieved from http://hdl.handle.net/2022/21622

PERVASIVE TECHNOLOGY INSTITUTE
INDIANA UNIVERSITY

RESEARCH TECHNOLOGIES
INDIANA UNIVERSITY
University Information Technology Services
Pervasive Technology Institute

In late 2014, The National Center for Genome Analysis Support (NCGAS) sent a short institutional review board (IRB) approved survey to ~5000 researchers who now have, or have had, NSF funding in the biological sciences. To ease IRB approval, this was a single-mailing, with no reminders. Thus, an overwhelming response wasn't expected. We received 53 responses, however the responses returned were detailed and thoughtful. From the answers, we concluded that a minority were already NCGAS uses. The majority are the audience we had hoped to contact: nation-wide biology researchers engaged in genome science.

This document includes the IRB-approved contact letter, the survey itself, and pie-charts that summarize the results of multi-choice questions. Attached also is a spread sheet that has all the results of the survey. The spreadsheet allows one to see the set of answers for each researcher, and includes the answers to essay/write-in questions (columns marked in blue).

The respondents covered a broad range of research methods (Fig. 1), from RNA sequencing (RNAseq), to genomics, to proteomics. Interestingly the most common approach was RNAseq, an area in which NCGAS has specialized. Researchers used a range of next generation sequencing (NGS) centers (a write in) and the majority reported that the center they used did not provide analysis, or they gave more complicated write-in answers, denoted as 'both/other' in Fig. 2. Only 21% gave an unequivocal 'yes'. 'Both/other' included centers that provided only initial analysis, ones where the center was too expensive to use, and researchers who simply chose to do their own. Clearly, sequencing centers are not providing researchers with the bioinformatics support they desire.

Researchers obtained their applications from various sources. The most common approach was to identify promising applications in the literature, and install it locally themselves (Fig. 3). The second most common are researchers who wroite their own. Only 17% reported using national resources such as NCGAS, Galaxy, and iPlant.

When asked which national resources researchers knew of, and then used, the two charts (Fig. 4 and 5) are nearly identical. Researchers knew what they used, and used what they knew. The four resources that dominated the answers were, in decreasing order: NCBI, the Broad Institute (an NCGAS partner), Galaxy (another area of NCGAS investment), and XSEDE. Only 8% reported using iPlant, and 10% CIPRES.

Finally, we asked what needs researchers had or foresaw (Fig. 6), and what services would prove most helpful to them (Fig. 7). While grant-supported bioinformatics personnel is the largest perceived need (Fig. 6), there was quite an even distribution across categories. Personnel was followed closely by time, but these two might be seen as a given in academic research. These were followed by short-term data storage, long term data storage, data transfer, and CPUs/memory, all with roughly equal scores. When offered a range of possible helpful services (Fig. 7), the first answer was curated published applications, followed closely by bioinformaticians-on-call and Galaxy-available applications; CPU, memory, and reference data were also common answers.

The write-in answers generally support the above conclusions. A few representative or informative responses are included here:

• "The speed of nr BLASTp of large transcriptome data sets (~5K contigs) takes about two months on our equipment (46 processors, 256g RAM). A high performance cluster dedicated to large BLAST jobs (i.e. no wall time limitations for working jobs) would be a huge help."

• "My grant cannot support a bioinformatician or experienced postdoc. So I'm left to do analysis on my own. So basic CS [computer science] issues are often obstacles."

• "Penn State is my home institution and main sequencing center. We have Galaxy and Biostars, but these are general tool and information site for DIY [do it yourself] approaches. They differ from the more extensive analyses that IU has been providing directly to its users."

• "We have had 3 CSPs from Joint Genome Institute. JGI is great and produces high quality sequence. They also have great informaticians, however their IMG [Integrated Microbial Genomes supports the annotation, analysis and distribution of microbial genome and metagenome data sets] is clunky, glitchy and limited in capability. Each implementation is different, so that if my colleagues have an IMG account set up at a different time they don't necessarily have the same functionality. It is awkward sharing data on the IMG system and requires intervention by IMG to give data access to colleagues. Assembly and annotation at JGI have been done by various facilities at different times, sometimes at JGI, sometimes at Oak Ridge and sometimes at Los Alamos. Although they re very responsive to input regarding assembly and annotation, it is unsatisfying to rely on second hand input only on how these processes are done. We also use commercial sequencing operations (e.g. Beckman Coulter Genomics) and have access to other resources through collaborators."

• "I don't know what/where our bioinformatics collaborator does the computation. I am hoping to do a sabbatical so that I can know more about bioinformatics, so that I might be able to answer questions such as these, or maybe even be able to handle the data, myself. I'm quite sorry to have to confess that I don't know enough to be able to answer your survey intelligently."

• "Yes; It is expensive to keep up with licensing fees for commercial products. Local installations of freeware are unsupported and tend to require special local knowledge, which is often lost when the staff member or student who installed it leaves the lab. Online resources come and go and it is difficult to know which provide superior capabilities. It is impossible to keep up with literature that compare different methods so we end up choosing based on convenience, chance and hear say."

• "The biggest impediment to discovery by biologists is the need to rely on others with knowledge of impenetrable systems and obscure acronyms to process and interpret data. Don't know how to fix this, but on some level user friendly platforms programs like Geneious more

than make up for their lack of power by providing an intuitive platform that encourages free exploration and experimentation with data."

In conclusion, the expressed needs of our researcher respondents are closely matched by the areas that NCGAS has emphasized in its original proposal and in the development since. We believe the questions were general enough that we can be confident that NCGAS services are needs felt by a national audience. To generalize, researchers need bioinformaticians to help them through their analysis; readily available stable applications, both as line-command and Galaxy-wrapped versions, and ready access to HPC resources.

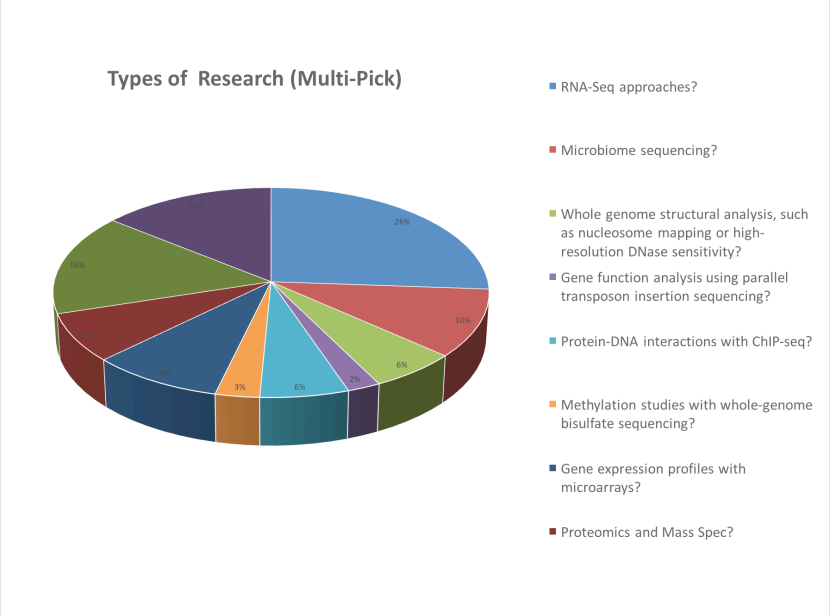

Figure 1

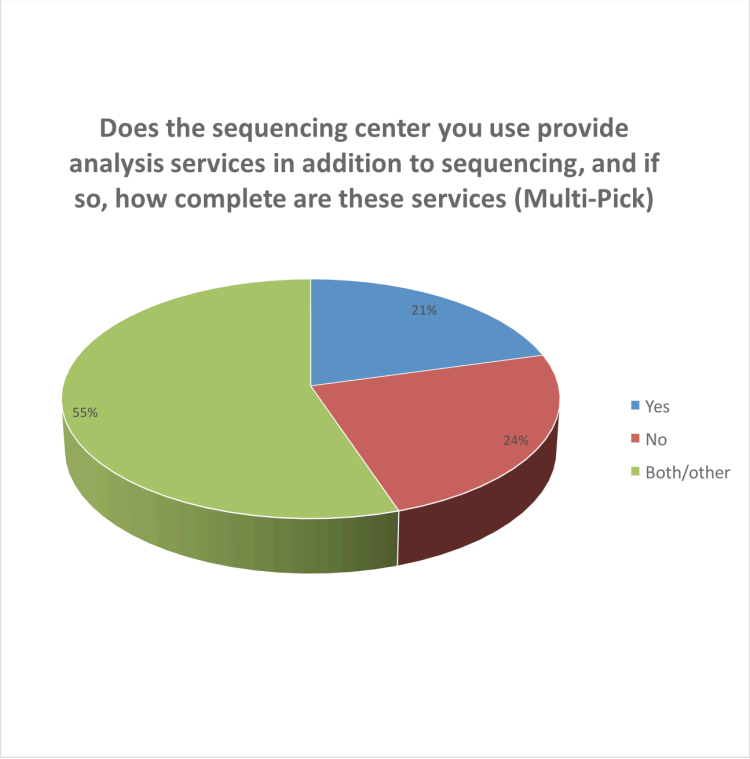

Figure 2

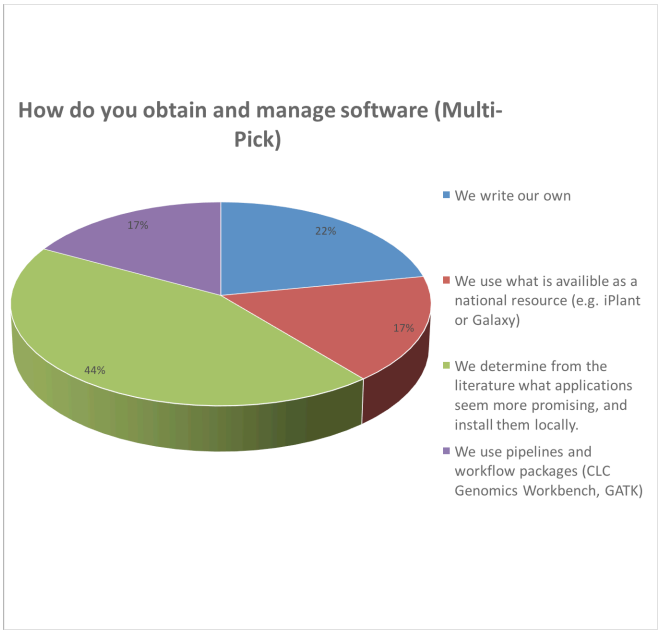

Figure 3

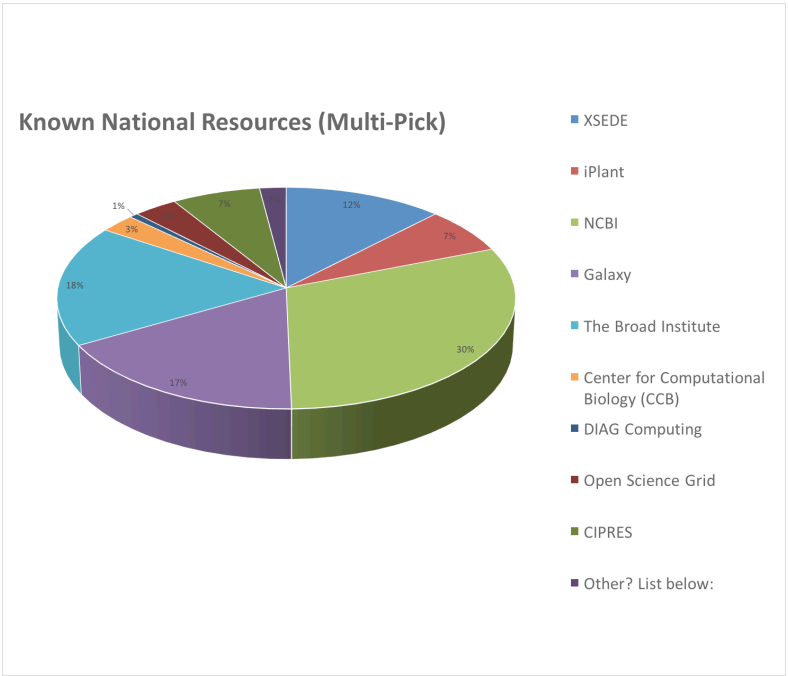

Figure 4

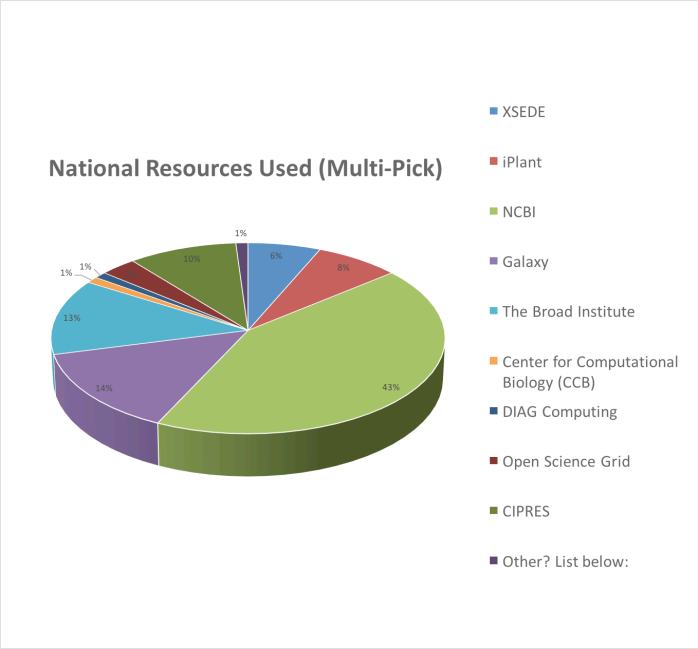

Figure 5

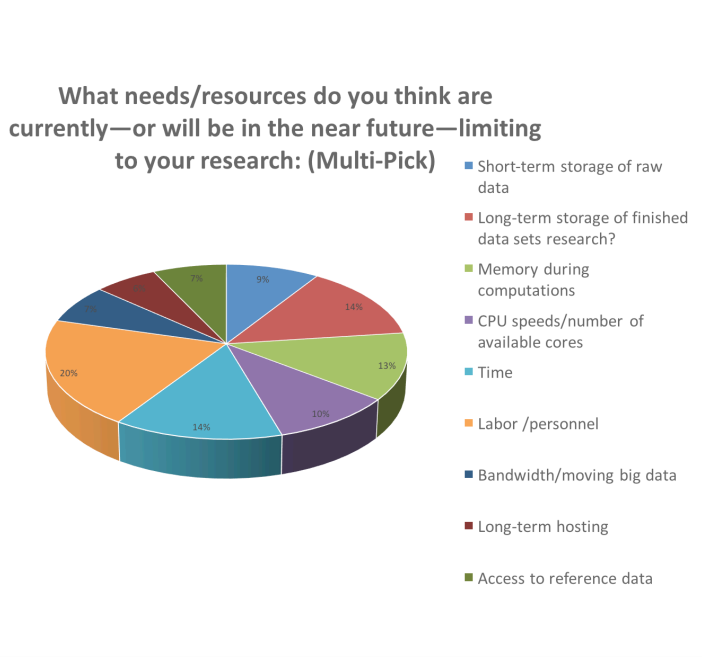

Figure 6

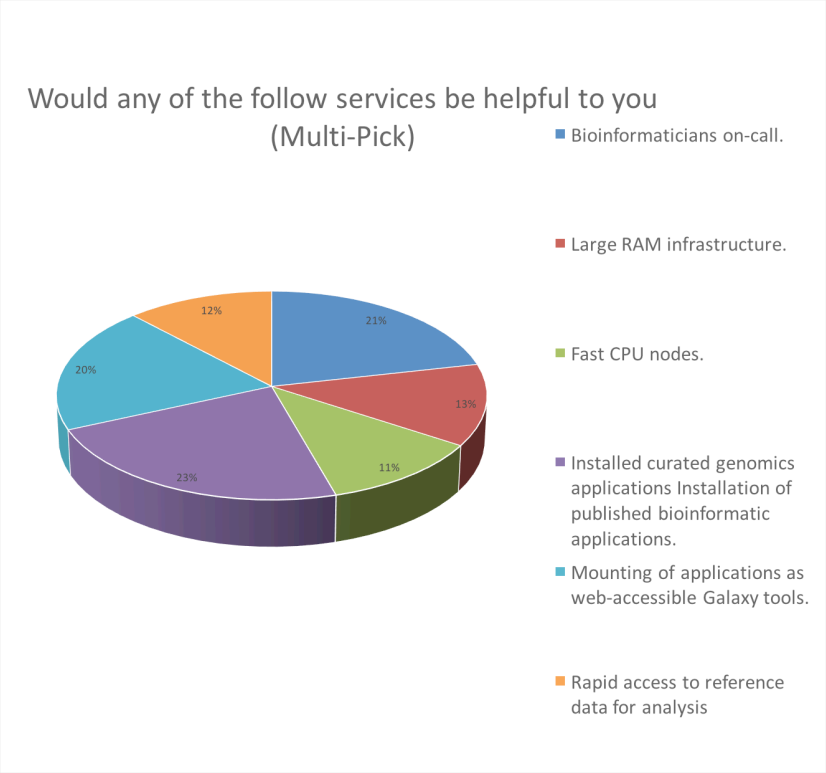
Would any of the follow services be helpful to you (Multi-Pick)
Bioinformaticians on-call.
Large RAM infrastructure.
Fast CPU nodes.
Installed curated genomics applications Installation of published bioinformatic applications.
Mounting of applications as web-accessible Galaxy tools.
Rapid access to reference data for analysis
21%
13%
11%
23%
20%
12%

Figure 7

Initial cover letter:

Dear Bioscience Researcher,

The National Center for Genome Analysis Support (NCGAS) at Indiana University is an initiative to provide NSF-funded life scientists with access and support to specialized computational resources on the national cyberinfrastructure. The goal is to create a single virtual system that lowers the computational barriers for biologists, bioinformaticians, physician-scientists, life-science students, and anyone, such as yourself, who is conducting research that uses next-generation DNA sequencing.

Toward this goal, we are conducting a survey in an effort to ascertain the needs of bioscience researchers in the general field of genomics (including transcriptomics, metagenomics, etc.), as we look forward to the growth of NCGAS services. As we aim to provide services that best match the current and future needs of the community, especially NSF-funded life scientists, we ask that you take a few moments to share with us your thoughts in this regard. The survey consists of less than a dozen questions and should take not more than 10 minutes to complete. Your feedback will be used to improve and expand services to the user community and to aid in the decision-making processes related to resource allocation.

The survey can be accessed here: https://redcap.uits.iu.edu/surveys/?s=yaUXW6DDIj

The Indiana Clinical and Translational Sciences Institutes REDCap system administers the survey and assures that your responses will remain completely confidential. Neither your name nor your organization will be associated with your responses.

If you have any questions about this survey or how the results will be used, please feel free to contact Tom Doak, NCGAS Manager, Indiana University, at tdoak@iu.edu, or (812) 856-0115.

Sincerely,

-Craig

Craig Stewart, Ph.D.
Primary Investigator, NCGAS
Executive Director, PTI Associate Dean Research Technologies

The National Center for Genome Analysis Support is administering this questionnaire on its own behalf. If you have any difficulties or questions about the study, please e-mail help@ncgas.org for assistance. If you do not wish to participate, please simply disregard this message.

Confidential

NSF PI Survey

NCGAS is an NSF-funded national center to aid researchers in managing, analyzing, and presenting research based on high-throughput sequencing. Our original focus was to provide the computational strength and curated applications needed for de novo genome assembly. Since then, we have been involved in developing tools for RNA-Seq projects in a wide range of organisms. These services are free for NSF-funded projects, and we currently work with investigators from across the country. In our efforts to improve our services to meet your changing needs, we would like to ask you about what your (NSF-associated life science researchers) needs are.

Does your research involve genomics, transcriptomics, microbiomes, or other high-throughput based sequencing approaches (generally NGS methods)?

If so, we would greatly appreciate your answers to a few more questions. If not, please return to your coffee with our thanks.

Does your research involve:

- ☐ RNA-Seq approaches?
- ☐ Microbiome sequencing?
- ☐ Whole genome structural analysis, such as nucleosome mapping or high-resolution DNase sensitivity?
- ☐ Gene function analysis using parallel transposon insertion sequencing?
- ☐ Protein-DNA interactions with ChIP-seq?
- ☐ Methylation studies with whole-genome bisulfate sequencing?
- ☐ Gene expression profiles with microarrays?
- ☐ Proteomics and Mass Spec?
- ☐ Genome annotation and ortholog discovery?
- ☐ Other? Please list.

Other research focus: _____

What sequencing centers have you used in the past? Can you describe good features and/or changes you would like to see in these centers' services? _____

Does the sequencing center you use provide analysis services in addition to sequencing, and if so, how complete are these services? _____

If your lab or collaborators perform some or all of the analysis, do you use computational hardware that is:

- ☐ Within your own lab?
- ☐ Provided by your organization/institution
- ☐ Provided by a national resource, such as XSEDE, iPlant, NCGAS, Galaxy Main.
- ☐ Other? Describe your resources below.

Briefly describe the hardware/ compute resources you use: _____

If your lab or collaborators perform some or all of the analysis, how do you obtain and manage software? Check all that apply.

- ☐ We write our own
- ☐ We use what is availible as a national resource (e.g. iPlant or Galaxy)
- ☐ We determine from the literature what applications seem more promising, and install them locally.
- ☐ We use pipelines and workflow packages (CLC Genomics Workbench, GATK)

Is installing and maintaining software ever an impediment, given changing versions, exotic dependencies, etc.?

- ☐ Yes
- ☐ No

Please describe any issues with software maintenance: _____

PERVASIVE TECHNOLOGY INSTITUTE

INDIANA UNIVERSITY

RESEARCH TECHNOLOGIES

INDIANA UNIVERSITY

University Information Technology Services
Pervasive Technology Institute

Please check all the national resources that you know about:

- ☐ iPlant
- ☐ XSEDE
- ☐ NCBI
- ☐ Galaxy
- ☐ The Broad Institute
- ☐ Center for Computational Biology (CCB)
- ☐ DIAG Computing
- ☐ Open Science Grid
- ☐ CIPRES
- ☐ Other? List below:

Other national resources known: _____

Please check all the national resources that you have used:

- ☐ iPlant
- ☐ XSEDE
- ☐ NCBI
- ☐ Galaxy
- ☐ The Broad Institute
- ☐ Center for Computational Biology (CCB)
- ☐ DIAG Computing
- ☐ Open Science Grid
- ☐ CIPRES
- ☐ Other? List below:

Other national resources used: _____

What needs/resources do you think are currently - or will be in the near future - limiting to your research?

- ☐ Short-term storage of raw data
- ☐ Long-term storage of finished data sets
- ☐ Memory during computations
- ☐ CPU speeds/number of available cores
- ☐ Time
- ☐ Labor /personnel
- ☐ Bandwidth/moving big data
- ☐ Long-term hosting
- ☐ Access to reference data

Would any of the follow services be helpful to you?

- ☐ Bioinformaticians on-call.
- ☐ Large RAM infrastructure.
- ☐ Fast CPU nodes.
- ☐ Installed curated genomics applications
- ☐ Installation of published bioinformatic applications.
- ☐ Mounting of applications as web-accessible Galaxy tools.
- ☐ Rapid access to reference data for analysis
- ☐ Other? Please describe:

Other resources that would be of help to you: _____